Journal of Statistical
Theory & Methods
1996, Vol. 7, pp 87~92

The Distributions of Variance Components in Two Stage Regression Môdel

Dong Joon Park[1]

Abstract A regression model with nested erroe structure is considered. The regression model includes two error terms that are independent and normally distributed with zero means and constant variances. This error structure of the model gives correlated response variables. The distributions of variance components in the regression model with nested error structure are dervied by using theorems for quadratic forms.

Keywords : mixed model, variance component

1. Introduction

For many years regression model has been one of the most heavily-used techniques in statistics. It has been applied in many different areas including social sciences, physical and biological sciences, business and technology, and humanities. It has been used to describe the relationship between the response and predictor variables.

Exact representation of regression model are not possible because of random errors associated with factors not included in the model. In classical regression model, these errors are assumed to be uncorrelated and normally distributed with zero mean and constant variance. This article considers the multiple regression model where the responses are correlated. In particluar, we consider two stage regrssion model, i.e., the multiple regression model with one-fold nested error structure. This model could be regarded as a multi-factor covariance model with multiple concomitant variables. This model is appropriate to the data collected using two stage cluster designs. This model includes two error terms. One is assiciated with the first-stage sampling unit and the other with the second-stage sampling unit. These two error terms are independent and normally distributed

[1] Department of Applied Mathematics National Fisheries University of Pusan, Pusan, 608-737, Korea

with zero means and constant variances. However, this error structure gives correlated response variables.

This article derives the distributions of variance components associated with the first and second stage sampling unit. Park and Burdick(1993) proposed the confidence intervals on the variance components in simple regression model with one-fold nested error structure. Aitken and Longford(1986) showed ignoring the nesting structure is not appropriate to estimate regression coefficients. Tsubaki et al.(1995) proposed the methods to estimate regression coefficients.

2. Two stage regression model

The two stage regression model is written as

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 X_{k_1 1} + \cdots + \beta_{p1} X_{k_{p1} p1} + \delta_i \\ &\quad + \gamma_1 X_{ij1} + \cdots + \gamma_{p2} X_{ijp2} + \varepsilon_{ij} \\ k_1 &= 1, \cdots, n_1; \ \cdots \ k_{p1} = 1, \cdots, n_{p1} \\ i &= 1, \cdots, l_1; \qquad j = 1, \cdots, l_2 \end{aligned} \tag{2.1}$$

where Y_{ij} is the jth observation in the ith cell(group), β_0 is an intercept term, $\beta_1, \cdots, \beta_{p1}$ are unknown parameters associated with primary units, $X_{k_1 1}, \cdots, X_{k_{p1} p1}$ are fixed predictor variables in the primary unit, $\gamma_1, \cdots, \gamma_{p2}$ are unknown parameters associated with secondary units, $X_{ij1}, \cdots, X_{ijp2}$ are fixed predictor variables in the secondary unit, δ_i is a random error term in the primary unit, ε_{ij} is a random error term in the secondary unit, δ_i and ε_{ij} are jointly independent normal random variables with zero means and variances σ_δ^2 and σ_ε^2, respectively. The index l_1 is the number of different combinations(cells) of levels among X_{ij}'s, i.e., $l_1 = n_1 \times n_2 \times \cdots \times n_{p1}$ and l_2 is the number of repetitions within an ith cell. We consider the balanced case where l_2's are same for all i's. Since β's, γ's, X_{ij}'s, and X_{ijk}'s are fixed, and δ_i and ε_{ij} are random, model (2.1) is a mixed model.

The model (2.1) is written in matrix notation,

$$\underline{Y} = ZX_1\underline{\beta} + X_2\underline{\gamma} + Z\underline{\delta} + \underline{\varepsilon} \tag{2.2.1}$$

$$= Z\underline{U} + X_2\underline{\gamma} + \underline{\varepsilon} \tag{2.2.2}$$

$$= X\underline{\alpha} + \underline{\xi} \tag{2.2.3}$$

where

$$\underline{U} = X_1\underline{\beta} + \underline{\delta}, \ X = (ZX_1 \ \ X_2), \ \underline{\alpha} = \left(\begin{smallmatrix} \beta \\ \underline{\gamma} \end{smallmatrix}\right), \text{ and } \underline{\xi} = Z\underline{\delta} + \underline{\varepsilon},$$

where $\underline{Y}$ is an $l_1 l_2 \times 1$ vector of observations, Z is an $l_1 l_2 \times l_1$ design matrix with 0's and 1's, i.e., $Z = \oplus_{i=1}^{l_1} \underline{1}_{l_2}$ where $\underline{1}_{l_2}$ is an $l_2 \times 1$ column vector of 1's and $\oplus$ is the direct sum operator, X_1 is an $l_1 \times (p_1+1)$ matrix of known values with a column of 1's in the first column and p_1 columns of X_{ij}'s from the second column to the p_1th column, $\underline{\beta}$ is a $(p_1+1)\times 1$ vector of parameters associated with X_{ij}'s, X_2 is an $l_1 l_2 \times p_2$ matrix of known values with p_2 columns of X_{ijk}'s, from the first column to the p_2th column, $\underline{\gamma}$ is a $p_2 \times 1$ vector of parameters associated with X_{ijk}'s, $\underline{\delta}$ is an $l_1 \times 1$ vector of random error terms, and $\underline{\varepsilon}$ is an $l_1 l_2 \times 1$ vector of random error terms. In particular,

$$\underline{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1l_2} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2l_2} \\ \vdots \\ Y_{l_1 1} \\ Y_{l_1 2} \\ \vdots \\ Y_{l_1 l_2} \end{pmatrix}, \quad Z = \begin{pmatrix} 10\cdots0 \\ 10\cdots0 \\ \vdots \\ 10\cdots0 \\ 01\cdots0 \\ 01\cdots0 \\ \vdots \\ 01\cdots0 \\ \vdots \\ 00\cdots1 \\ 00\cdots1 \\ \vdots \\ 00\cdots1 \end{pmatrix}, \quad X_2 = \begin{pmatrix} X_{111} & X_{112} & \cdots & X_{11p_2} \\ X_{121} & X_{122} & \cdots & X_{12p_2} \\ & \vdots & & \\ X_{1l_2 1} & X_{1l_2 2} & \cdots & X_{1l_2 p_2} \\ X_{211} & X_{212} & \cdots & X_{21p_2} \\ X_{221} & X_{222} & \cdots & X_{22p_2} \\ & \vdots & & \\ X_{2l_2 1} & X_{2l_2 2} & \cdots & X_{2l_2 p_2} \\ & \vdots & & \\ X_{l_1 11} & X_{l_1 12} & \cdots & X_{l_1 1p_2} \\ X_{l_1 21} & X_{l_1 22} & \cdots & X_{l_1 2p_2} \\ & \vdots & & \\ X_{l_1 l_2 1} & X_{l_1 l_2 2} & \cdots & X_{l_1 l_2 p_2} \end{pmatrix}, \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1l_2} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2l_2} \\ \vdots \\ \varepsilon_{l_1 1} \\ \varepsilon_{l_1 2} \\ \vdots \\ \varepsilon_{l_1 l_2} \end{pmatrix},$$

$$X_1 = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p_1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p_1} \\ & & \vdots & & \\ 1 & X_{n_1 1} & X_{n_1 2} & \cdots & X_{n_{p_1} p_1} \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p_1} \end{pmatrix}, \quad \underline{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{p_2} \end{pmatrix}, \quad \underline{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{l_1} \end{pmatrix}.$$

From (2.2.3), the variance-covariance matrix of $\underline{Y}$ is

$$Var(\underline{Y}) = \sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1 l_2}, \tag{2.3}$$

since $\underline{\delta} \sim N(\underline{0}, \sigma_\delta^2 I_{l_1})$ and $\underline{\varepsilon} = N(\underline{0}, \sigma_\varepsilon^2 I_{l_1 l_2})$ where I_{l_1} is an $l_1 \times l_1$ identity matrix. From the assumptions in (2.1) and equation (2.3),

$$\underline{Y} \sim N(X\underline{\alpha}, \sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1 l_2}). \tag{2.4}$$

The regression sums of squares of model (2.1) are now investigated. The reductions in sums of squares of the model are attributable to fitting the primary and secondary fixed variables and are expressed into the quadratic forms. Let G_1, G_2 and G_3 be generalized inverses of $X'X, X^{*'}X^{*}$, and $\overline{X}_2'\overline{X}_2$, respectively, where $X^{*} = (X_1\ X_2^{*})$, $X_2^{*} = \frac{Z'}{l_2}X_2$, $\overline{X}_2 = WX_2$, and $W = I_{l_1l_2} - ZZ'/l_2$. Define $H_1 = XG_1X'$, $H_2 = X^{*}G_2X^{*'}$ and $H_3 = \overline{X}_2G_3\overline{X}_2'$. Now consider the quadratic forms $R_1 = \underline{Y}'(I_{l_1l_2} - H_1)\underline{Y}$, $R_2 = \underline{Y}'\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z'}{l_2}\underline{Y}$, and $R_3 = \underline{Y}'W'(I_{l_1l_2} - H_3)W\underline{Y}$.

The quadratic form R_1 is obtained by regressing the response variable on the primary and secondary fixed variables in (2.2.3). The quadratic form R_2 is determined by computing the regression of $\overline{Y}_{i.}$ on X_{ij} and $\overline{X}_{i.k}$ where $\overline{Y}_{i.} = \sum_{j=1}^{l_2} Y_{ij}/l_2$ and $\overline{X}_{i.k} = \sum_{j=1}^{l_2} X_{ijk}/l_2$. The quadratic form R_3 is calculated by the regression of Y_{ij} on the secondary fixed variables, X_{ijk}.

3. The distributional properties of variance components

Theorem 3.1 Under the distributional assumptions in (2.1), the matrix $R_1/(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2})$ is chi-squared distributed with $l_1l_2 - p_1 - p_2 - 1$ degrees of freedom.

Proof. Consider quadratic form $R_1 = \underline{Y}'(I_{l_1l_2} - H_1)\underline{Y}$. Multiplying $(I_{l_1l_2} - H_1)/(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2})$ by (2.3) gives $I_{l_1l_2} - H_1$. Note that $(I_{l_1l_2} - H_1)$ is idempotent with $r(I_{l_1l_2} - H_1) = l_1l_2 - p_1 - p_2 - 1$. In addition, $(X\underline{\alpha})'(I_{l_1l_2} - H_1)(X\underline{\alpha}) = 0$. It follows by theorem 2(Searle 1971,p.57) that $R_1/(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2})$ is chi-squared distrbuted with $l_1l_2 - p_1 - p_2 - 1$ degrees of freedom.

Theorem 3.2 Under the distributional assumptions in (2.1), the quadratic form $R_2/(\sigma_\delta^2 + \sigma_\varepsilon^2/l_2)$ is a chi-squared random variable with $l_1 - p_1 - p_2 - 1$ degrees of freedom.

Proof. Consider quadratic form $R_2/(\sigma_\delta^2 + \sigma_\varepsilon^2/l_2)$. Multiplying $\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z'}{l_2}/(\sigma_\delta^2 + \sigma_\varepsilon^2/l_2)$ by (2.3) gives

$$\frac{\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z'}{l_2}}{(\sigma_\delta^2 + \sigma_\varepsilon^2/l_2)}(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2}) = \frac{Z}{l_2}(I_{l_1} - H_2)Z'$$

since $Z'Z/l_2 = I_{l_1}$. In addition, $\frac{Z}{l_2}(I_{l_1} - H_2)Z'$ is idempotent. Since $H_2X_1 = X_1, H_2X_2^{*} = X_2^{*}$, and $H_2Z'X_2 = Z'X_2$,

$$(X\underline{\alpha})'\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z}{l_2}(X\underline{\alpha}) = 0$$

and $r(\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z'}{l_2}) = l_1 - p_1 - p_2 - 1$ degrees of freedom.

Theorem 3.3 Under the distributional assumptions in (2.1), the quadratic form $R_3 / \sigma_\varepsilon^2$ is a chi-squared random variable with $l_1 l_2 - l_1 - p_2$ degrees of freedom.
Proof. Consider quadratic form $R_3 / \sigma_\varepsilon^2$. Since W and H_3 are symmetric and idempotent, and $WH_3 = H_3 W = H_3$, the matrix $W\text{-}H_3$ is idempotent with $r(W\text{-}H_3) = l_1 l_2 - l_1 - p_2$. Note that

$$\begin{aligned}\frac{W'(I_{l_1l_2} - H_3)W}{\sigma_\varepsilon^2}(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2}) &= \frac{(W - H_3)}{\sigma_\varepsilon^2}(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2}) \\ &= W - H_3\end{aligned}$$

since $WZ = H_3 Z = 0_{l_1l_2 \times l_1}$ where $0_{l_1l_2 \times l_1}$ is an $l_1 l_2 \times l_1$ matrix of zeros. Note that

$$\begin{aligned}X'W'(I_{l_1l_2} - H_3)WX &= X'WX - X'H_3X \\ &= \begin{pmatrix} X_1'Z'WZX_1 & X_1'Z'WX_2 \\ X_2'WZX_1 & X_2'WX_2 \end{pmatrix} - \begin{pmatrix} X_1'Z'H_3ZX_1 & X_1'Z'H_3X_2 \\ X_2'H_3ZX_1 & X_2'H_3X_2 \end{pmatrix} \\ &= \begin{pmatrix} 0_{(p_1+1)\times(p_1+1)} & 0_{(p_1+1)\times p_2} \\ 0_{p_2\times(p_1+1)} & X_2'WX_2 \end{pmatrix} - \begin{pmatrix} 0_{(p_1+1)\times(p_1+1)} & 0_{(p_1+1)\times p_2} \\ 0_{p_2\times(p_1+1)} & X_2'H_3X_2 \end{pmatrix} \\ &= 0_{(p_1+p_2+1)\times(p_1+p_2+1)}\end{aligned}$$

since $WZ = H_3 Z = 0_{l_1l_2 \times l_1}$, $Z'W = Z'H_3 = 0_{l_1 \times l_1 l_2}$ and $H_3 X_2 = \overline{X}_2$. Furthermore, $r(W'(I_{l_1l_2} - H_3)W) = l_1 l_2 - l_1 - p_2$. Hence, by theorem 2(Searle 1971, p.57), $R_3 / \sigma_\varepsilon^2$ is a chi-squared random variable with $l_1 l_2 - l_1 - p_2$ degrees of freedom.

Theorem 3.4 Under the distributional assumptions in (2.1), the quadratic forms $R_2 / (\sigma_\delta^2 + \sigma_\varepsilon^2 / l_2)$ and $R_3 / \sigma_\varepsilon^2$ are independent.
Proof. Note that

$$\frac{Z}{l_2}(I_{l_1} - H_2)\frac{Z'}{l_2}(\sigma_\delta^2 ZZ' + \sigma_\varepsilon^2 I_{l_1l_2})W'(I_{l_1l_2} - H_3)W = 0_{l_1l_2 \times l_1l_2}$$

by using that $Z'W' = 0_{l_1 \times l_1 l_2}$. Hence, by theorem 4(Searle 1971, p.59), the two quadratic forms $R_2 / (\sigma_\delta^2 + \sigma_\varepsilon^2 / l_2)$ and $R_3 / \sigma_\varepsilon^2$ are independent.

4. Conclusions

This article presents the distributions of variance components in two stage regression model. The distributions of variance components are found by using the quadratic forms. Three quadratic forms are based on the regression sums of squares computed by response variable on different combinations of fixed predictor variables. The distributions of variance components in the model can be used to find confidence intervals on variance components for future research.

References

Aitken, M. and Longford, N. T. (1986), *Statistical modelling issues in school effectiveness studies(with discussion)*, Journal of Royal Statistical Society A., 14 A, 1-43.

Park, D. J. and Burdick, R. K. (1993), *Confidence intervals on the among group variance component in a simple linear regression model with a balanced one-fold nested error structure*, Communications in Statistics-Theory and Methods, 22(12), 3435-3452.

Searle, S. R. (1971), *Linear Models*, New York: John Wiley & Sons inc.

Tsubaki, M., Tsubaki, H, and Kusumi, M. (1995), *A new estimation method for a useful class of mixed models*, Proceedings of the 9th Asia Quality Management Symposium - Quality Enhancement for Global Prosperity, 275-280.